Projet de thèse: Evaluation et développement de méthodes statistiques intégrant les annotations fonctionnelles dans la prédiction de trait complexes pour l'optimisation des schémas de sélection

Mots-clés : Prédiction génomique, annotations fonctionnelles, modèles linéaires mixtes, modèles bayésiens, intelligence artificielle, simulation, génétique quantitative, sélection variétale

Laboratoire d'accueil : UMR AGAP Institut, Avenue Agropolis - 34398 Montpellier Cedex 5 (https://umr-agap.cirad.fr/

Encadrants :
Directeur de thèse : David Pot (CIRAD, HDR), Généticien en appui aux programmes de sélection
Encadrant de thèse : Vincent Garin (CIRAD), Biostatisticien

Contexte et problématique de l'étude

La sélection de variétés de plantes adaptées aux évolutions du climat et des systèmes agricoles futurs est un des enjeux majeurs de la recherche agronomique actuelle. La prédiction de caractères complexes dans des organismes biologiques comme les animaux ou les plantes à l'aide d'information génétique, connue sous le nom de prédiction génomique, est une application importante des statistiques pour améliorer l'efficacité de la sélection (Hickey et al., 2017; Xu et al., 2020). Fortement ancrée dans la théorie des modèles mixtes, la prédiction génomique s'est enrichie d'apports des statistiques Bayésiennes ainsi que plus récemment des algorithmes issus de l'apprentissage machine (Machine Learning, ML).
La compréhension de la construction et de la variabilité des phénotypes cibles de la sélection implique deux grands types d'équipes de recherche :

- des équipes de généticiens quantitatifs et de biostatisticiens tentant d'expliquer la variation génétique des caractères par des modèles statistiques ancrés sur la variabilité nucléotidique (Hickey et al., 2017; Xu et al., 2020)
- des équipes de biologistes et physiologistes moléculaires visant quant à elles l'identification des mécanismes physiologiques et des gènes et réseaux moléculaires sous-jacents (Brooks et al., 2021; Davidson, 2010) permettant d'expliquer la construction des phénotypes finaux et leurs expressions.

Si la prédiction génomique commence à prendre en compte les informations issues des études d'identification de gènes majeurs (Xiaogang Liu et al., 2019; Xu et al., 2020), elle ne tire pas encore complètement parti de la compréhension fine des mécanismes moléculaires impliqués dans la construction des phénotypes. L'intégration des informations génétiques et biologiques issues des approches multi-omiques dans les modèles de prédiction génomique est pourtant une stratégie prometteuse (Azodi et al., 2020; Boyle et al., 2017; Chateigner et al., 2020; de las Heras-Saldana et al., 2020; Gage et al., 2021; Giri et al., 2021; Hu et al., 2019; Xuanyao Liu et al., 2019; Schrag et al., 2018; Westhues et al., 2019, 2017; Zhengcao, 2019). Néanmoins cette approche basée sur des caractérisations multi-omiques (génomique, transcriptomique, protéomique, métabolomique…) des différents candidats à la sélection se traduit par des coûts relativement élevés encore peu compatibles avec les contraintes financières des programmes de sélection. Dans ce contexte, la mobilisation d'annotations fonctionnelles déjà disponibles dans la littérature comme des Go-Terms (Edwards et al., 2016) ou des propriétés positionnelles ou évolutives des polymorphismes considérés (Ramstein et al., 2020; Ramstein and Buckler, 2022) peut permettre une importante réduction des coûts.

Plusieurs méthodes statistiques ont été proposées pour intégrer les annotations fonctionnelles dans les modèles de prédiction génomique, avec des approches allant des modèles mixtes utilisant une pré-sélection de marqueurs (Edwards et al., 2016) à des approches bayésiennes permettant une formulation plus complexe des effets a priori (Moser et al. 2015 ; Mollandin et al., 2022; Zheng et al., 2024). La comparaison de ces méthodes ainsi que l'évaluation d'approches de machine learning ont encore été très peu explorées. En matière de sélection, l'importance de l'interaction génotype par environnement et la nécessité de traiter des données produites dans plusieurs environnements sont incontournables (Malosetti et al. 2013). L'adaptation de méthodes de prédictions intégrant la connaissance fonctionnelle a priori au contexte de prédiction multi-environnemental constitue un challenge statistique nécessitant la prise en compte d'effets génétiques variables. Finalement, l'intégration de ces nouveaux outils en appui des programmes de sélection constitue un objectif d'application concret des méthodes qui seront proposées dans ce projet de thèse.

Objectifs du projet de thèse

L'objectif général de ce projet de thèse est d'explorer les bénéfices liés à l'ajout d'annotations fonctionnelles dans les approches de prédictions génomiques mises en œuvre dans différents plans de croisements et différentes structures de schéma de sélection d'espèces tropicales et méditerranéennes.

Ce projet s'articulera autour de 4 axes.

- Axe 1 : Evaluation par simulation des différentes méthodes actuellement disponibles dans le contexte de plans de croisement, de populations et de structures de schéma de sélection contrastants
- Axe 2 : Développement et évaluation de nouvelles méthodes d'intégration des informations fonctionnelles notamment au travers des approches de Machine Learning
- Axe 3 : Exploration de la pertinence de l'intégration des informations fonctionnelles dans le contexte de caractères soumis à des interactions génotype x environnement
- Axe 4 : Application des méthodes d'intégration des annotations fonctionnelles dans la prédiction génomique sur des jeux de données réels correspondant à différents plans de croisements, populations et schémas de sélection

Matériels et Méthodes

Les 3 premiers axes de la thèse seront basés sur des approches de simulation qui s'ancreront sur des plans de croisements et des schémas de sélection ayant déjà été analysés ou décrits dans le cadre des travaux de l'équipe Génétique et Innovation Variétale de l'UMR AGAP Institut. Dans un premier temps, ces travaux de simulation seront initiés sur la base d'un large dispositif de croisement multiparental qui a été développé en partenariat entre le CIRAD, l'Institut d'Economie Rural (IER) du Mali et l'Institut de recherche international sur les espèces de zones semi-arides tropicales (ICRISAT). Ce dispositif de type BCNAM (BackCross Nested Association Mapping, (Jordan et al., 2011)) comprenant plus de 3900 familles a été évalué au sein de plusieurs environnements et il a servi de base au développement d'une méthodologie innovante de détection de QTL et de leurs interactions avec l'environnement (Garin et al., 2024). En complément de ce dispositif, un nombre limité de types de plans de croisements, de populations et de schémas de sélection représentatifs des approches mobilisées dans le cadre de la sélection des plantes autogames sera aussi considéré pour tester l'intérêt de ces approches d'intégration en fonction des contextes d'utilisation.

Sur ces différents dispositifs, les approches déjà existantes d'intégration des annotations fonctionnelles au sein de la prédiction génomique seront testées (Axe 1). Ces approches se déclinent actuellement en deux grands types : des approches basées sur des modèles linéaires mixtes (Edwards et al., 2016; Ramstein et al., 2020; Speed and Balding, 2014) et des approches bayésiennes (Lloyd-Jones et al., 2019; Mollandin et al., 2022; Zeng et al., 2023; Zheng et al., 2024).

En complément, la pertinence des approches de Machine Learning dans le contexte de l'ajout d'annotations fonctionnelles sera aussi explorée (Axe 2). Actuellement les bénéfices des approches de Machine Learning à la prédiction génomique sans ajout d'informations fonctionnelles sont encore débattus. En effet, des résultats contrastants ont été observés notamment en lien avec la taille des jeux d'entrainements utilisés ainsi que des capacités de paramétrage de ces modèles (Jubair and Domaratzki, 2023; Lourenço et al., 2024). Ces approches seront intéressantes à explorer pour tirer parti des larges corpus d'annotations fonctionnelles disponibles.
Un des défis actuels de la sélection des plantes est de gérer les interactions génotype x environnement qui amènent à une modification de l'ordre des candidats à la sélection en fonction de leurs environnements de culture. Des couplages de la sélection génomique avec des informations environnementales ont été proposées récemment. L'utilité de l'ajout des annotations fonctionnelles dans ce contexte seront aussi explorées en considérant différents niveaux de GXE et les évolutions de l'architecture génétique des caractères cibles en fonction des conditions environnementales (Axe 3).
Sur la base des simulations réalisées (Axes 1 à 3), les méthodes sélectionnées ou développées, les plus adaptées aux différents contextes seront mises en œuvre sur des jeux de données réels. Les annotations fonctionnelles qui seront mobilisées correspondront aux QTL déjà détectés soit dans les dispositifs ciblés soit au sein d'analyses externes, aux positions des polymorphismes par rapport aux gènes à proximité, aux annotations des gènes contenant les polymorphismes (classes fonctionnelles, signature de sélection, appartenance à des modules d'expression).

Résultats attendus

- Axe 1 : Evaluation par simulation des différentes méthodes actuellement disponibles dans le contexte de plans de croisement, de populations et de structures de schéma de sélection contrastants

Au sein de cet axe, les méthodes basées sur des modèles linéaires mixtes ainsi que sur des approches bayésiennes déjà développées chez les plantes, les animaux et l'humain seront testées sur des plans de croisements et des schémas de sélection représentatifs de la sélection des espèces autogames. Ce premier axe permettra d'identifier les limites et les intérêts des différentes méthodes en fonction des architectures génétiques des caractères, de la structure des dispositifs de croisements et des qualités des annotations fonctionnelles.

- Axe 2 : Développement et évaluation de nouvelles méthodes d'intégration des informations fonctionnelles notamment au travers des approches de Machine Learning

Des approches de Machine Learning ont été proposées pour optimiser l'efficacité de la prédiction génomique notamment dans le cas de la prise en compte des informations environnementales. En revanche, malgré l'intérêt évident de ces approches, elles n'ont été pour l'instant que rarement mobilisées pour prendre en compte les annotations fonctionnelles qui sont largement disponibles. Il est attendu de cet axe de caractériser les intérêts des approches de Machine Learning dans ce contexte en proposant des méthodes d'exploration des architectures des approches de Machine Learning (les CNN en particulier) et de définition des hyperparamètres sous-jacents.

- Axe 3 : Exploration de la pertinence de l'intégration des informations fonctionnelles dans le contexte de caractères soumis à des interactions génotype x environnement

L'intérêts de l'intégration des annotations fonctionnelles en fonction des intensités et de la structure des interactions GXE sera évalué à travers des approches de simulations.

- Axe 4 : Application des méthodes d'intégration des annotations fonctionnelles dans la prédiction génomique sur des jeux de données réels correspondant à différents plans de croisements, populations et schémas de sélection

Sur la base des résultats obtenus au travers des approches de simulation, les approches les plus pertinentes seront mises en œuvre sur des jeux de données réels correspondant à différents plans de

croisements, différentes populations et structures de schémas de sélection. Ces stratégies seront entre autres mobilisées dans le cadre de 3 projets en cours basés sur le sorgho ciblant i) son adaptation à l'intensification des périodes de stress post floral (projet ANR SorDrought), ii) l'optimisation de sa qualité du grain pour l'alimentation avicole et humaine (Projet CASDAR Semences NItropSorgh) et iii) la compréhension des bases génétiques des interactions plantes-sol dans un contexte d'optimisation de la séquestration du carbone dans le sol (Projet PEPR FairCarbon RhizoSeqC). Ces données réelles basées sur des dispositifs multi parentaux de différentes ampleurs (ANR SorDrought et PEPR Fair Carbon) et des populations non-apparentées (NitroSorg) offriront une vision globale de l'intérêt des intégrations des annotations fonctionnelles.
Ces analyses pourront être complétées par des jeux de données disponibles sur le riz notamment.

Conditions scientifiques matérielles (conditions de sécurité spécifiques) et financières du projet de recherche :
Ce projet de thèse est financé dans le cadre du projet ANR SorDrought « Caractérisation de nouveaux traits physiologiques pour aider l'amélioration de la tolérance au stress hydrique post-floral chez le sorgho ») et du projet PEPR FairCarbon RhizoSeqC « Optimiser la rhizodéposition pour augmenter la séquestration de carbone dans les agrosystèmes » qui ont tous deux pour objectifs de développer des méthodologies d'appui à la sélection dans le cadre de partenariats privés- publics.
Ce projet sera mené au sein de l'UMR AGAP Institut et plus spécifiquement au sein de l'équipe Génétique Innovation Variétale qui cible l'optimisation des méthodologies de sélection au service des programmes d'amélioration végétale. En complément, l'UMR AGAP Institut développe actuellement son expertise sur la mobilisation des approches de Machine Learning dans le cadre de la prédiction génomique et le candidat sélectionné bénéficiera de cette dynamique. Enfin, le candidat sélectionné sera aussi intégré au groupe GQMS² (Génétique Quantitative, Méthodologie de la Sélection et Sélection) qui correspond à un groupe d'animation transversal de l'UMR.
Le candidat intègrera enfin l'UMR dans un contexte où plusieurs étudiants en thèse développent différentes approches d'optimisation de la sélection génomique et phénomique. Des synergies pourront donc être mise en place en termes de construction de synthèses bibliographiques et de liens entre les différents projets de thèse (impacts des structures des dispositifs sur les capacités de prédiction, impacts des architectures génétiques…)

Ouverture Internationale :

La participation du (de la) doctorant(e) à plusieurs conférences internationales lui permettra de développer son réseau international. Le doctorant sera aussi amené à partager ses avancées et questionnements avec des partenaires du Sud qui ont été impliqués dans l'acquisition des données pour les différentes espèces tropicales mobilisées et qui sont fortement intéressés par la mobilisation des informations moléculaires au sein de leur programme de sélection. Des financements seront recherchés pour permettre un séjour de 6 mois dans une équipe développant des travaux sur les méthodologies statistiques pertinentes dans un contexte d'intégration d'annotation fonctionnelles à l'évaluation des valeurs génétiques (cadre de la génétique humaine, animale ou végétale).

Collaborations envisagées :

Au niveau de l'UMR AGAP institut, des collaborations seront mises en place avec les équipes Phenomen (Phénotypage et modélisation des plantes dans leur environnement agro-climatique) et GSP (Génome et sélection des pérennes) qui se mobilisent actuellement dans l'utilisation des approches de Deep Learning (CNN notamment) dans la prédiction génomique.

Au niveau national, le / la doctorante sera impliqué(e) dans le réseau NetBio qui est un réseau méthodologique en Mathématiques appliquées qui ciblait initialement les aspects de construction de réseaux d'interaction biologiques mais qui s'étend actuellement vers l'intégration de ces données dans un contexte de mobilisation de la variabilité génétique au service des applications en sélection.

En complément, comme mentionné dans le cadre du paragraphe précédent dédié à l'ouverture internationale, des financements seront recherchés pour permettre une mobilité de la personne sélectionné au sein d'une équipe internationale travaillant sur les aspects statistiques.

Objectifs de valorisation des travaux de recherche du doctorant : diffusion, publication et confidentialité, droit à la propriété intellectuelle *:
Les résultats issus des travaux de la thèse, seront valorisés sous forme de publications dans des revues à Facteur d'impacts et de présentation dans des colloques internationaux (Eucarpia Biometrics, Plant and Animal Genome, Statgen : Conference on Statistics in Genomics and Genetics). L'intérêt de travailler sur l'optimisation d'une méthodologie très récente est la capacité à valoriser relativement facilement les résultats de recherche, qui sont très attendus sur les questions spécifiques que nous souhaitons explorer au cours de ce projet de thèse.

Profil recherché

- Master Recherche en statistiques, mathématiques appliquées, bio-informatique, ou autre domaine avec un fort accent sur les méthodes quantitatives, en particulier les modèles mixtes et/ou les approches bayésiennes
- Fort intérêt pour le test et le développement de méthodologies statistiques
- Intérêt pour les applications dans le domaine de l'amélioration des plantes avec un fort attrait pour les biostatistiques et la bio-informatique
- Utilisation d'un langage d'analyse et de programmation (R, Python, C++)
- Attrait pour le travail en équipe incluant notamment des généticiens, des sélectionneurs et des statisticiens
- Le désir de travailler dans un contexte international lie au développement de l'agriculture est un plus.

Formation continue du doctorant

Les formations nécessaires pour la réalisation du travail de thèse seront choisies parmi les catalogues offerts par l'Université de Montpellier (Ecole Doctorale GAIA), AGAP Institut, le Cirad et l'INRAE selon les besoins. En termes généraux, l'étudiant suivra les formations sur la déontologie et l'intégrité scientifique dans les métiers de la recherche ainsi que sur la rédaction de publications, la présentation des résultats scientifiques et la vulgarisation. En termes plus spécifiques au projet proposé seront *a minima* ciblées : les nouveaux outils et permettant de mettre en œuvre les méthodologies de Deep learning

Suivi d'avancement du projet de thèse

Trois comités de suivi de thèse seront organisés (1/an), pour assurer le bon déroulement de la thèse et éventuellement recadrer/réorienter certaines parties si besoin. En cas de problème, des comités supplémentaires pourront être organisés pour les résoudre. En plus du représentant de la filière doctorale qui permettra de satisfaire au mieux aux exigences de l'ED et d'un représentant d'AGAP Institut, les autres membres du comité seront choisis, dans le respect des règles, pour apporter des compétences dans les disciplines des biostatistiques et de la génétique quantitative.

Informations administratives :

Année universitaire de 1ère inscription en doctorat*: 2024
Date de début de la thèse*: 01/10/2024
Date limite de candidature*: 23h59 05/07/2024

Contacts :

David Pot (HDR) : Généticien en appui aux programmes de sélection
CIRAD Agricultural Research for Development
Biological Systems Department
AGAP Mixed Research Unit "Genetic Improvement and Adaptation of Mediterranean and Tropical Plants"
Genetics and Varietal Innovation Team
Building 3 - Office 129
TA A-108 / 03 - Avenue Agropolis - 34398 Montpellier Cedex 5
France
E-mail: david.pot@cirad.fr
Téléphone : +33 6 51 75 13 76

Vincent Garin : Biostatisticien en appui aux programmes de sélection
CIRAD Agricultural Research for Development
Biological Systems Department
AGAP Mixed Research Unit "Genetic Improvement and Adaptation of Mediterranean and Tropical Plants"
Genetics and Varietal Innovation Team
Building 3bis – Office 152
TA A-108 / 03 - Avenue Agropolis - 34398 Montpellier Cedex 5
France
E-mail: vincent.garin@cirad.fr

Publications pertinentes de l'équipe (les personnes de l'équipe sont indiquées en gras) :

Garin, V., Diallo, C., Tékété, M.L., Théra, K., Guitton, B., Dagno, K., Diallo, A.G., Kouressy, M., Leiser, W., Rattunde, F., Sissoko, I., Touré, A., Nébié, B., Samaké, M., Kholovà, J., Frouin, J., **Pot, D., Vaksmann, M**., Weltzien, E., Témé, N., Rami, J.-F., 2024. Characterization of adaptation mechanisms in sorghum using a multireference back-cross nested association mapping design and envirotyping. Genetics iyae003. https://doi.org/10.1093/genetics/iyae003

Burgarella, C., Berger, A., Glémin, S., David, J., Terrier, N., Deu, M., **Pot, D**., 2021. The Road to Sorghum Domestication: Evidence From Nucleotide Diversity and Gene Expression Patterns. Front. Plant Sci. 12, 1706. https://doi.org/10.3389/fpls.2021.666075

Garin, V., Choudhary, S., Murugesan, T., Kaliamoorthy, S., Diancumba, M., Hajjarpoor, A., Chellapilla, T.S., Gupta, S.K., Kholovà, J., 2023a. Characterization of the Pearl Millet Cultivation Environments in India: Status and Perspectives Enabled by Expanded Data Analytics and Digital Tools. Agronomy 13, 1607. https://doi.org/10.3390/agronomy13061607

Garin, V., Diallo, C., Tekete, M.L., Thera, K., Guitton, B., Dagno, K., Diallo, A.G., Kouressy, M., Leiser, W., Rattunde, F., Sissoko, I., Toure, A., Nebie, B., Samake, M., Kholova, J., **Frouin, J., Pot, D.**, **Vaksmann, M**., Weltzien, E., Teme, N., Rami, J.-F., 2023b. Characterization of adaptation mechanisms in sorghum using a multi-reference back-cross nested association mapping design and envirotyping. https://doi.org/10.1101/2023.03.11.532173

Garin, V., Malosetti, M., van Eeuwijk, F., 2020a. The usefulness of multi-parent multi-environment QTL analysis: an illustration in different NAM populations (preprint). Genetics. https://doi.org/10.1101/2020.02.03.931626

Garin, V., Wimmer, V., Borchardt, D., Malosetti, M., van Eeuwijk, F., 2020b. The influence of QTL allelic diversity on QTL detection in multi-parent populations: a simulation study in sugar beet (preprint). Genetics. https://doi.org/10.1101/2020.02.04.930677

Garin, V., Wimmer, V., Mezmouk, S., Malosetti, M., van Eeuwijk, F., 2017. How do the type of QTL effect and the form of the residual term influence QTL detection in multi-parent populations? A case study in the maize EU-NAM population. TAG Theor. Appl. Genet. Theor. Angew. Genet. 130, 1753–1764. https://doi.org/10.1007/s00122-017-2923-3

Hennet, L., Berger, A., Trabanco, N., Ricciuti, E., Dufayard, J.-F., Bocs, S., Bastianelli, D., Bonnal, L., Roques, S., Rossini, L., Luquet, D., Terrier, N., **Pot, D**., 2020. Transcriptional Regulation of Sorghum Stem Composition: Key Players Identified Through Co-expression Gene Network and Comparative Genomics Analyses. Front. Plant Sci. 11. https://doi.org/10.3389/fpls.2020.00224

Nguyen, V.H., Morantte, R.I.Z., Lopena, V., Verdeprado, H., Murori, R., Ndayiragije, A., Katiyar, S.K., Islam, M.R., Juma, R.U., Flandez-Galvez, H., Glaszmann, J.-C., Cobb, J.N., **Bartholomé, J**., 2023. Multi-environment Genomic Selection in Rice Elite Breeding Lines. Rice 16, 7. https://doi.org/10.1186/s12284-023-00623-6

Références Bibliographiques

Azodi, C.B., Pardo, J., VanBuren, R., Campos, G. de los, Shiu, S.-H., 2020. Transcriptome-Based Prediction of Complex Traits in Maize. Plant Cell 32, 139–151. https://doi.org/10.1105/tpc.19.00332

Boyle, E.A., Li, Y.I., Pritchard, J.K., 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169, 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Brooks, M.D., Juang, C.-L., Katari, M.S., Alvarez, J.M., Pasquino, A., Shih, H.-J., Huang, J., Shanks, C., Cirrone, J., Coruzzi, G.M., 2021. ConnecTF: A platform to integrate transcription factor–gene interactions and validate regulatory networks. Plant Physiol. 185, 49–66. https://doi.org/10.1093/plphys/kiaa012

Chateigner, A., Lesage-Descauses, M.-C., Rogier, O., Jorge, V., Leplé, J.-C., Brunaud, V., Roux, C.P.-L., Soubigou-Taconnat, L., Martin-Magniette, M.-L., Sanchez, L., Segura, V., 2020. Gene expression predictions and networks in natural populations supports the omnigenic theory. BMC Genomics 21, 1–16. https://doi.org/10.1186/s12864-020-06809-2

Davidson, E.H., 2010. Emerging properties of animal gene regulatory networks. Nature 468, 911–920. https://doi.org/10.1038/nature09645

de las Heras-Saldana, S., Lopez, B.I., Moghaddar, N., Park, W., Park, J., Chung, K.Y., Lim, D., Lee, S.H., Shin, D., van der Werf, J.H.J., 2020. Use of gene expression and whole-genome sequence information to improve the accuracy of genomic prediction for carcass traits in Hanwoo cattle. Genet. Sel. Evol. 52, 54. https://doi.org/10.1186/s12711-020-00574-2

Edwards, S.M., Sørensen, I.F., Sarup, P., Mackay, T.F.C., Sørensen, P., 2016. Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster. Genetics 203, 1871–1883. https://doi.org/10.1534/genetics.116.187161

Gage, J.L., Mali, S., McLoughlin, F., Khaipho-Burch, M., Monier, B., Bailey-Serres, J., Vierstra, R.D., Buckler, E.S., 2021. Variation in upstream open reading frames contributes to allelic diversity in protein abundance. https://doi.org/10.1101/2021.05.25.445499

Garin, V., Diallo, C., Tekete, M.L., Thera, K., Guitton, B., Dagno, K., Diallo, A.G., Kouressy, M., Leiser, W., Rattunde, F., Sissoko, I., Toure, A., Nebie, B., Samake, M., Kholova, J., Frouin, J., Pot, D., Vaksmann, M., Weltzien, E., Teme, N., Rami, J.-F., 2023. Characterization of adaptation mechanisms in sorghum using a multi-reference back-cross nested association mapping design and envirotyping. https://doi.org/10.1101/2023.03.11.532173

Giri, A., Khaipho-Burch, M., Buckler, E.S., Ramstein, G.P., 2021. Haplotype associated RNA expression (HARE) improves prediction of complex traits in maize. PLOS Genet. 17, e1009568. https://doi.org/10.1371/journal.pgen.1009568

Hickey, J.M., Chiurugwi, T., Mackay, I., Powell, W., 2017. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat. Genet. 49, 1297–1303. https://doi.org/10.1038/ng.3920

Hu, X., Xie, W., Wu, C., Xu, S., 2019. A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol. J. 17, 2011–2020. https://doi.org/10.1111/pbi.13117

Jordan, D.R., Mace, E.S., Cruickshank, A.W., Hunt, C.H., Henzell, R.G., 2011. Exploring and Exploiting Genetic Variation from Unadapted Sorghum Germplasm in a Breeding Program. Crop Sci. 51, 1444–1457. https://doi.org/10.2135/cropsci2010.06.0326

Jubair, S., Domaratzki, M., 2023. Crop genomic selection with deep learning and environmental data: A survey. Front. Artif. Intell. 5. https://doi.org/10.3389/frai.2022.1040295

Liu, Xuanyao, Li, Y.I., Pritchard, J.K., 2019. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177, 1022-1034.e6. https://doi.org/10.1016/j.cell.2019.04.014

Liu, Xiaogang, Wang, H., Hu, X., Li, K., Liu, Z., Wu, Y., Huang, C., 2019. Improving Genomic Selection With Quantitative Trait Loci and Nonadditive Effects Revealed by Empirical Evidence in Maize. Front. Plant Sci. 10, 1129. https://doi.org/10.3389/fpls.2019.01129

Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., Metspalu, A., Wray, N.R., Goddard, M.E., Yang, J., Visscher, P.M., 2019. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. 10, 5086. https://doi.org/10.1038/s41467-019-12653-0

Lourenço, V.M., Ogutu, J.O., Rodrigues, R.A.P., Posekany, A., Piepho, H.-P., 2024. Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. BMC Genomics 25, 152. https://doi.org/10.1186/s12864-023-09933-x

Malosetti, M., Ribaut, J.-M., and van Eeuwijk, F. A., 2013. The statistical analysis of multi-environment data: modeling genotype-by environment interaction and its genetic basis. Frontiers in physiology, 4:44

Mollandin, F., Gilbert, H., Croiseau, P., Rau, A., 2022. Accounting for overlapping annotations in genomic prediction models of complex traits. BMC Bioinformatics 23, 365. https://doi.org/10.1186/s12859-022-04914-5

Ramstein, G.P., Buckler, E.S., 2022. Prediction of evolutionary constraint by genomic annotations improves functional prioritization of genomic variants in maize. Genome Biol. 23, 183. https://doi.org/10.1186/s13059-022-02747-2

Ramstein, G.P., Larsson, S.J., Cook, J.P., Edwards, J.W., Ersoz, E.S., Flint-Garcia, S., Gardner, C.A., Holland, J.B., Lorenz, A.J., McMullen, M.D., Millard, M.J., Rocheford, T.R., Tuinstra, M.R., Bradbury, P.J., Buckler, E.S., Romay, M.C., 2020. Dominance Effects and Functional Enrichments Improve Prediction of Agronomic Traits in Hybrid Maize. Genetics 215, 215–230. https://doi.org/10.1534/genetics.120.303025

Schrag, T.A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., Melchinger, A.E., 2018. Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. Genetics 208, 1373–1385. https://doi.org/10.1534/genetics.117.300374

Speed, D., Balding, D.J., 2014. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 24, 1550–1557. https://doi.org/10.1101/gr.169375.113

Westhues, M., Heuer, C., Thaller, G., Fernando, R., Melchinger, A.E., 2019. Efficient genetic value prediction using incomplete omics data. TAG Theor. Appl. Genet. Theor. Angew. Genet. 132, 1211–1222. https://doi.org/10.1007/s00122-018-03273-1

Westhues, M., Schrag, T.A., Heuer, C., Thaller, G., Utz, H.F., Schipprack, W., Thiemann, A., Seifert, F., Ehret, A., Schlereth, A., Stitt, M., Nikoloski, Z., Willmitzer, L., Schön, C.C., Scholten, S., Melchinger, A.E., 2017. Omics-based hybrid prediction in maize. TAG Theor. Appl. Genet. Theor. Angew. Genet. 130, 1927–1939. https://doi.org/10.1007/s00122-017-2934-0

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., Prasanna, B.M., Olsen, M.S., Wang, G., Zhang, A., 2020. Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. Plant Commun. 1, 100005. https://doi.org/10.1016/j.xplc.2019.100005

Zeng, J., Zheng, Z., Liu, S., Sidorenko, J., Yengo, L., Turley, P., Ani, A., Wang, R., Nolte, I., Snieder, H., Yang, J., Wray, N., Goddard, M., Visscher, P., 2023. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex trauts within and between ancestries. Eur. Neuropsychopharmacol., Abstracts of the World Congress of Psychiatric Genetics (WCPG), October 10-14, 2023 75, S29–S30. https://doi.org/10.1016/j.euroneuro.2023.08.063

Zheng, Z., Liu, S., Sidorenko, J., Wang, Y., Lin, T., Yengo, L., Turley, P., Ani, A., Wang, R., Nolte, I.M., Snieder, H., Yang, J., Wray, N.R., Goddard, M.E., Visscher, P.M., Zeng, J., 2024. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Nat. Genet. 56, 767–777. https://doi.org/10.1038/s41588-024-01704-y

Zhengcao, L., 2019. Integrating Omics Data into Genomic Prediction. Georg-August-University Göttingen, Göttingen.